

PR #51432 完整报告

vllm-project/vllm

[Bugfix][multi_modal] Fix pos_ids being uninitialized for minicpmv2.6 in hf runner

合并时间: 2026-08-08 12:09

原文链接: <http://prhub.com.cn/vllm-project/vllm/pull/51432>

执行摘要

- 一句话: 修复 MiniCPM-V2.6 HF 测试中 pos_embed 未初始化导致的 flaky
- 推荐动作: 值得快速精读, 尤其适合维护多模态测试与 HF runner 的工程师。亮点在于: 一是用最小化 instrumentation 将 flaky 问题精确定位到 transformers 加载阶段; 二是用 assert restored 保证 workaround 在未来模型变化时快速失效而不是静默出错; 三是明确“上游修复后 revert”的临时补丁策略, 值得借鉴。

功能与动机

PR body 指出 #48413 在 Multi-Modal Models (Extended Generation 3) TG 引入了两类失败, 一类是 NameError: name 'List' is not defined, 另一类是 AttributeError: 'MiniCPMV' object has no attribute 'all_tied_weights_keys'; 后者修复后测试仍 flaky, 表现为 HF 输出全 NaN 或 hf/vllm 输出不一致。通过临时 instrumentation 定位到 NaN 出现在模型加载阶段而非前向过程: vLLM 的 pos_embed 是正常 sin/cos 表, HF 的却是全 0。根因是 transformers 的 _move_missing_keys_from_meta_to_device 把 non-persistent 缓冲区替换为 torch.empty_like, 而 _initialize_missing_keys 只重建 rotary embedding 缓冲区, 不恢复 pos_embed。

实现拆解

1. 根因定位: 作者用 standalone 脚本对比 vLLM 与 HF 加载后的 resampler.pos_embed, 并用 _set_2d_pos_cache 重跑验证, 确认问题发生在 transformers 权重加载阶段, 而非前向计算。
2. 新增 helper: 在 tests/models/multimodal/generation/vlm_utils/model_utils.py 新增 _restore_resampler_pos_cache, 遍历 hf_model.model.modules(), 对具备 _set_2d_pos_cache 的模块以 module.max_size 和 module.pos_embed.device 重算位置表, 并用 assert restored 兜底, 防止模型结构调整后静默失效。
3. 接入 patch 流程: 在 minicpmv_26_patch_hf_runner 中, 于 patch generate 之前调用 _restore_resampler_pos_cache, 确保 HF 参考解码使用正确的位置表。
4. 验证: 6 个用例 (3 个 multi-image + 3 个 single-image) 循环 20 次共 120 次执行全部通过; 无配置文件、schema 或生产代码改动。

关键文件:

- tests/models/multimodal/generation/vlm_utils/model_utils.py (模块 测试工具; 类别 test; 类型 test-coverage; 符号 _restore_resampler_pos_cache): 唯一变更文件。新增 _restore_resampler_pos_cache 并在 minicpmv_26_patch_hf_runner 中调用, 修复 HF 加载时 pos_embed 被置为未初始化内存导致的测试 flaky。

关键符号: _restore_resampler_pos_cache, minicpmv_26_patch_hf_runner

关键源码片段

tests/models/multimodal/generation/vlm_utils/model_utils.py

唯一变更文件。新增 _restore_resampler_pos_cache 并在 minicpmv_26_patch_hf_runner 中调用, 修复 HF 加载时 pos_embed 被置为未初始化内存导致的测试 flaky。

```
def _restore_resampler_pos_cache(hf_model: HfRunner) -> None:
    """Recompute the resampler's 2D sin/cos position cache after loading.

    `from_pretrained` materializes non-persistent buffers as uninitialized
    memory, leaving `pos_embed` as zeros or NaN instead of the values computed
    in `__init__`.
    """
    restored = 0
    for module in hf_model.model.modules():
        if not hasattr(module, "_set_2d_pos_cache"):
            continue
        # 在 buffer 当前所在设备上重算位置缓存, 等价于重新执行 __init__ 里的初始化逻辑
        module._set_2d_pos_cache(module.max_size, module.pos_embed.device)
        restored += 1
    # 若模型结构变化导致找不到 resampler, 立即失败而不是静默通过
    assert restored, "no resampler pos cache found to restore"

def minicpmv_26_patch_hf_runner(hf_model: HfRunner) -> HfRunner:
    # 先恢复被 transformers 破坏的 pos cache, 再 patch generate
    _restore_resampler_pos_cache(hf_model)

    orig_generate = hf_model.model.generate

    def _generate(self, *args, image_sizes=None, **kwargs):
        return orig_generate(*args, decode_text=False, **kwargs)

    hf_model.model.generate = types.MethodType(_generate, hf_model.model)

    return hf_model
```

评论区精华

review 讨论不多且没有代码评审注释。核心讨论是维护者 DarkLight1337 提出的策略:

“Let's fix CI first and revert once the issue has been fixed upstream”, 即把本改动视为针对 transformers 缺陷的临时 workaround, 待上游修复后回滚。AndreasKaratzas 审核后批

准 (LGTM) 。claude[bot] 提示该 PR 来自 fork, 自动 review 被禁用, 需维护者手动触发。

- 上游修复后的临时策略 (design): 先合并修复 CI, 等待上游 transformers 修复后 revert 本测试补丁。
- 维护者批准 (other): PR 获得维护者批准并合并。
- Fork PR 的自动 review 被禁用 (other): 未触发自动 review, 改由人工 review 完成。

风险与影响

- 风险: 本次变更全部落在测试代码中, 生产代码零接触, 回归风险极低。但该 workaround 依赖 transformers 对 non-persistent buffer 的处理行为, transformers 版本升级后可能不再需要或不再匹配; assert restored 能保证在模型结构调整或 _set_2d_pos_cache 消失时快速失败, 避免静默通过。若 MiniCPM-V 远程代码签名变化、或在非 CUDA 设备上执行, helper 中 .device 传递逻辑需保持正确。整体属于低风险、高可观测性的测试侧修复。
- 影响: 影响范围限定在多模态测试的 HF runner 参考路径, 特别是 minicpmv_26 的 6 个生成用例; 修复后 Extended Generation 3 TG 不再因 HF 全 NaN 或 hf/vllm 不一致而误报。对 vLLM 推理、配置、性能无任何影响。对团队而言, 该 PR 提供了一个应对第三方库加载缺陷的测试侧补丁模式, 并与 #51427 共同恢复同一 TG 的 CI 稳定性。
- 风险标记: 纯测试路径改动, 依赖 transformers 内部行为, 上游修复后需 revert

关联脉络

- PR #51427 [Bugfix][models_multimodal] Remote HF python code misses importing class: 同一 MiniCPM-V 多模态生成 TG 的另一半修复: 解决 PR body 中提到的 NameError: name 'List' is not defined (远程代码 List 未导入), 与本 PR 共同恢复该 TG 的 CI 稳定性。